AI RESEARCH SHOULD SCALE INSIGHT, NOT JUST COMPUTE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling compute has driven substantial progress in AI, but performance gains do not necessarily bring a corresponding increase in reusable knowledge about learning. **We argue that AI research should scale insight, not just compute**. We define insight as reusable, communicable knowledge with a stated scope of applicability. Systematically producing, testing, and accumulating such knowledge could deepen our understanding of learning and guide more effective training, supporting both the science of AI and AI-for-AI research. We explore this approach by actively running experiments and measuring observables to identify learning phenomena, formulate hypotheses, and test interventions. We present OPHIS (Observation–Problem–Hypothesis–Intervention–Speed-up), a conceptual framework for observation-guided research; ComfyResearch, a visual tool for constructing, reproducing, and sharing experiments; and the nanoGPT Observable Library, a dataset of observable trajectories recorded during training. Using the nanoGPT Observable Library, we observe attention-entropy patterns that differ across depth and warmup settings. Applying OPHIS to AutoResearch, we identify interventions that improve validation performance under a fixed training-step budget. We demonstrate ComfyResearch on grokking and edge-of-stability experiments and show how automated experiment generation supports the exploration of learning dynamics. We outline a research agenda for deepening human understanding of learning, accumulating and sharing insights, applying them in AutoResearch, and assessing their value under compute and time constraints.

1 INTRODUCTION

Early models such as support vector machines offered relatively tractable mathematical descriptions (Cortes & Vapnik, 1995). Deep learning shifted the focus toward scaling models, data, and compute (Kaplan et al., 2020; Sevilla et al., 2022). As AI systems have grown more capable, they have also become harder to understand (Simon et al., 2026). Historically, physics progressed from Tycho Brahe through Kepler to Newton (Cohen, 1985); chemistry progressed from alchemy through the periodic table to modern chemical theory (Brock, 2016). **Today's AI research remains largely in the first stage: extensive experimentation, but limited accumulation of reusable understanding**.

Within this largely empirical paradigm, scaling has become a central strategy for improving AI capabilities across the model lifecycle. During pre-training, researchers increase parameter counts, training tokens, and computational budgets (Brown et al., 2020; Kaplan et al., 2020). During post-training, scaling can involve larger collections of supervised or preference data, more reinforcement-learning rollouts, and greater optimization budgets (Ouyang et al., 2022; Kimi Team et al., 2025). At inference time, test-time scaling allocates additional computation to sampling, search, verification, or iterative refinement (Snell et al., 2025). Across these stages, scaling has delivered substantial empirical gains, and LLMs now serve as foundations for increasingly capable agentic systems (Brown et al., 2020; Ouyang et al., 2022; Yao et al., 2023).

Much of practical model development still relies on established heuristics, with new ideas tested through repeated computational experiments. Such experiments can tell us whether a modification improves performance, but not necessarily why it works or when the same idea should work elsewhere (Lipton & Steinhardt, 2018). Standard performance metrics can also obscure what happens during training. In small Transformers trained on modular addition, mechanistic analyses found

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

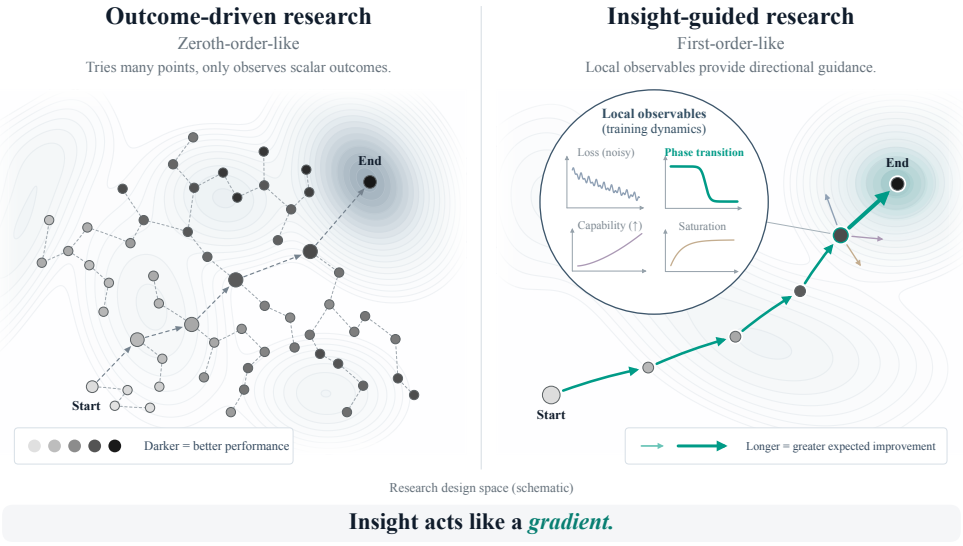


Figure 1: **Insight provides direction for research.** Outcome-driven research (left) compares experimental results to select what to try next. In insight-guided research (right), the enlarged node shows the training dynamics of three observables. These observations inform hypotheses that guide research directions, with matching curve and arrow colors. The teal arrow represents the direction with the greatest expected improvement and leads to the next experiment. The landscape, curves, and paths are schematic.

that the circuits responsible for generalization emerged well before the sharp improvement in test accuracy, revealing distinct stages of memorization, circuit formation, and cleanup (Nanda et al., 2023; Liu et al., 2022). Related work showed that models trained on the same task can implement qualitatively different internal algorithms, including the Clock and Pizza algorithms, depending on architecture, hyperparameters, and initialization (Zhong et al., 2023). Together, these results show that **the learning process can contain structure that is largely invisible from final performance alone**. When existing heuristics provide little guidance in a new setting, this missing information must instead be recovered through further experimentation. Observing learning dynamics during these experiments can help researchers decide what to try next.

Scaling has been an effective way to improve AI systems, but it does not by itself explain how or why learning succeeds. In other scientific and engineering fields, systematic experiments have often been used not only to improve individual systems, but also to identify principles that generalize across settings. For example, steam engines were already widely used when Carnot sought general principles governing their efficiency (Carnot, 1824). Similarly, Mendeleev's periodic law brought accumulated knowledge of the elements into a common framework and enabled predictions of the properties of elements not yet discovered (Mendeleev, 1889). We argue that AI experiments should also aim to produce such reusable knowledge. Understanding when and why a method works can reduce the need to rediscover effective choices through repeated trial and error in every new setting. Outcome-driven research resembles zeroth-order optimization, using observed performance to select what to try next. **Insights from learning dynamics can provide gradient-like guidance for this choice, making research more like first-order optimization** (Figure 1). In this sense, progress in AI should involve scaling insight as well as compute.

2 Insight can and should be scaled

In cognitive psychology, insight often refers to a sudden reinterpretation of a problem that leads to a non-obvious solution—an "aha moment" (Kounios & Beeman, 2014). This meaning emphasizes the moment of realization. **In this paper, we use insight to refer to knowledge about the system being studied**. Such knowledge may arise from an aha moment, but it can also develop through systematic

experiments or gradual reasoning. We therefore distinguish the knowledge itself from the process through which it is obtained, and give a more precise definition in the following subsection.

2.1 DEFINING INSIGHT

We use *insight* to refer to a reusable and communicable piece of knowledge with a stated scope of applicability. An insight goes beyond an isolated observation: within its scope, it should support qualitatively correct prediction or explanation, or help guide an intervention. This notion is closely related to accounts of scientific understanding that emphasize the ability to recognize characteristic behavior and apply knowledge in new situations (de Regt & Dieks, 2005; Krenn et al., 2022). An individual insight need not amount to a complete understanding of a system; it can instead constitute one component of such an understanding.

Communicability matters because scientific knowledge must be available for evaluation and reuse. Intuition that cannot be articulated is difficult for others to assess or apply. Recent reports describe mathematical contributions from GPT and Claude, including improved bounds on large prime gaps and on the proportion of zeros of the Riemann zeta function on the critical line (OpenAI, 2026; Anthropic, 2026). Yet producing results does not by itself ensure that the underlying ideas can be understood and reused by human researchers. This distinction is central to a recent declaration by 25 Fields Medalists, which emphasizes mathematical understanding and the transmission of ideas (Avila et al., 2026). Krenn et al. (2022) similarly propose successful transfer of understanding to a human expert as a sufficient criterion for an AI system to possess scientific understanding.

Insights can take many forms, including empirical regularities, governing relations, mechanistic explanations, constraints, and intervention principles. Importantly, **we do not require every insight to provide a complete mechanism**. A robust predictive relationship, supported by empirical or theoretical evidence, may already be scientifically useful. For example, Achille et al. (2019) found that the trace of the Fisher information matrix exhibits an early peak during training. In the settings they studied, this period also coincided with the greatest sensitivity to temporary image-blurring deficits. The correspondence reveals structure in the stages of learning even without a complete mechanistic account of why it occurs. Requiring a full mechanism before treating such a result as an insight would exclude much of the phenomenological knowledge from which deeper explanations are often developed.

Comparing and connecting local insights from toy models and controlled tasks may gradually build a broader understanding of learning. Such insights may also transfer beyond their original settings when an appropriate correspondence between settings is established. Under maximal update parametrization (μP), for example, several optimal hyperparameters remain approximately stable across model widths. μTransfer uses this relation to tune hyperparameters on smaller models and transfer them to larger ones, reducing the need for expensive large-model tuning (Yang et al., 2021).

2.2 SCIENCE OF AI

Recent work has emphasized the importance of understanding the fundamental principles underlying AI systems and developing a science of AI (Bartlett et al., 2021; Simon et al., 2026; Ferguson et al., 2026; Biderman et al., 2026). One challenge for such a science is that the metrics commonly used in model development provide only a limited view of the learning process. Loss curves and benchmark performance primarily characterize training outcomes, while learning can continue even when these metrics show little or no improvement (Barak et al., 2022; Mallinar et al., 2025; Kangaslahti et al., 2026). **This motivates studying training dynamics directly, which can reveal how models acquire features and how these features evolve over the course of training** (Biderman et al., 2026; Nanda et al., 2023). Such analyses can provide reusable insights into how models learn.

In Physics of AI, studying phenomena such as double descent (Nakkiran et al., 2020), grokking (Power et al., 2022), neural collapse (Papyan et al., 2020), and the edge of stability (Cohen et al., 2021) provides a way to develop such insights. Researchers propose and test hypotheses to explain when and why these phenomena occur. For example, Cohen et al. (2021) observed that, during full-batch gradient descent, the largest eigenvalue of the training loss Hessian hovers slightly above $2/\eta$, where η is the learning rate, while the loss continues to decrease over longer timescales. This raised the question of how training can make progress beyond the stability threshold predicted by

a local quadratic approximation. Damian et al. (2023) subsequently showed how motion in unstable directions can reduce curvature and restore stability. Their analysis yielded predictions for the loss and curvature during training, which they tested experimentally. **This example shows how investigating a phenomenon can lead to a testable explanation of learning dynamics**. Different phenomena expose different aspects of learning and provide clues to the principles behind them.

We argue that relying on incidental discoveries of new phenomena is too slow. LLM-based systems have achieved gold-medal-standard performance in olympiad mathematics and surpassed most human participants in competitive-programming evaluations (Luong & Lockhart, 2025; Guo et al., 2025). Yet predicting which internal algorithms models learn remains difficult beyond the controlled settings in which these algorithms have been studied (Zhong et al., 2023; Biderman et al., 2026). To accelerate the science of AI, **we need to scale phenomena**. This means systematically exploring a wider range of tasks, models, and training conditions to discover, reproduce, and characterize learning phenomena. Omnigrok illustrates how extending this exploration can improve understanding: it extended the study of grokking to image classification, sentiment analysis, and molecular property prediction under suitably chosen training conditions (Liu et al., 2023). These results showed that grokking is not specific to algorithmic datasets and supported a loss-landscape account of when it occurs.

2.3 Towards mechanistic AutoResearch

With tools and agent interfaces, LLMs can interact with external environments, execute actions, and use feedback to revise their behavior (Yao et al., 2023; Yang et al., 2024). In one demonstration, sixteen Claude agents, supported by a human-designed testing environment, produced a roughly 100,000-line C compiler in Rust (Carlini, 2026). These capabilities also support AI-for-AI research, where AI systems help develop and improve AI systems themselves. The AI Scientist connects idea generation, implementation, experimentation, analysis, and manuscript preparation within an automated research workflow (Lu et al., 2026). At Anthropic, agents developed and tested training methods that reduced ten types of alignment failures, with some methods also working on larger models (Chen et al., 2026).

However, research judgment remains a challenge. On TASTE, a benchmark of preferences over AI safety research proposals, the best evaluated model agreed with the benchmark labels on 60% of proposal pairs, compared with estimated human agreement of 77% (Baig et al., 2026).

We argue that a lack of insights that can guide research is a key bottleneck for automated AI research, particularly when designing training recipes. Human researchers draw on empirical rules to choose model components, configure training, and formulate controlled experiments. Both human and automated researchers need knowledge about how models learn, but developing a mature science of AI is a long-term undertaking. Making existing empirical rules available is, in our view, the most direct way to provide useful guidance in the meantime. We refer to practical, experimentally supported rules of this kind as *tricks*. When their scope of applicability is stated and their effects can be tested and reused, such rules constitute local insights in the sense defined above. To make automated AI research more effective, **we need to scale tricks**.

Existing training studies illustrate the practical effectiveness of such tricks. In large-batch ResNet-50 training, He et al. (2019) reported that adding learning-rate warmup increased ImageNet top-1 accuracy from 75.17% to 75.93%, and subsequently adding zero-gamma initialization increased it to 76.37%. In deep reinforcement learning, Engstrom et al. (2020) found that implementation-level optimizations accounted for much of PPO's performance advantage over TRPO in their experiments. These results show that training outcomes depend substantially on practical choices beyond a model architecture or an algorithm's core formulation. **Scaling tricks therefore means systematically collecting, testing, and sharing effective empirical rules, together with their conditions of applicability, so that both human and automated researchers can reuse them**.

3 Our exploration

Scaling insights is an ambitious goal, but we can begin by finding suitable ways to represent them. Code and tool calls use explicit syntax and structured representations. In procedural pretraining, structured data improved subsequent performance on algorithmic tasks, while shuffling tokens

within the pretraining sequences removed these gains (Jiang et al., 2026). **We argue that representing insights in a structured form can likewise help AI systems learn and reuse research knowledge**.

Researchers conduct experiments, such as ablation studies, to assess how a component affects model behavior. These experiments can yield observed regularities and experimentally supported hypotheses. A regularity observed under specified conditions can support expectations about what will happen in similar settings. Some of this understanding remains implicit in individual researchers' experience, making it difficult for others to access and reuse. In a performance-driven workflow, researchers instead focus on running experiments, ranking their results by loss or benchmark scores, and looking for patterns in those rankings.

AI agents can automate this workflow, from running experiments to ranking results and identifying patterns. However, automation alone does not resolve the limitations in research taste discussed in Section 2.3. One way to address these limitations is to express researchers' judgment as written guidance, such as skill files, that agents can use. We argue that a more basic limitation remains: when theory provides little guidance about which changes should work, choosing what to try next still depends heavily on guesses. **Automation can accelerate this trial-and-error process without necessarily making those choices better informed**.

Rather than waiting for learning phenomena to be discovered incidentally, we actively run experiments and measure a range of observables to explore how models learn. We allocate part of the computational budget to this systematic observation, using the findings to identify problems and formulate hypotheses (Figure 1, right). The remaining budget is used to test interventions suggested by these hypotheses, with the aim of accelerating training.

3.1 Observables

Physical laws often describe particular aspects of a system's behavior using a small number of relevant quantities. Newton's second law, for example, relates the net force on an object to the rate of change of its momentum. Although an object has many properties, including color, shape, and hardness, these properties do not explicitly enter this relation. An ideal gas contains a vast number of particles, yet its pressure, volume, particle number, and temperature are related by the simple equation $PV = Nk_BT$, where k_B is Boltzmann's constant. In the study of phase transitions, an order parameter, such as spontaneous magnetization in a ferromagnet, can distinguish different phases.

Studying learning dynamics likewise requires appropriate observables. In Omnigrok, weight norm provided a useful coordinate for analyzing reduced training and test loss landscapes, revealing a "Goldilocks" range associated with good generalization. This picture guided experiments that induced grokking on non-algorithmic tasks such as MNIST under suitably chosen training conditions (Liu et al., 2023). Learning can also suppress task-irrelevant variation while retaining useful features. For example, neural collapse shows how, during the terminal phase of classification training, last-layer features of training examples within each class converge toward their class mean while the class means remain separated (Papyan et al., 2020). Separately, training trajectories across a range of architectures and training configurations have been found to explore an effectively low-dimensional manifold in prediction space (Mao et al., 2024). These findings motivate our expectation that **regularities in learning dynamics can be described by a small number of key observables**.

A science of learning dynamics requires explicit state variables, rather than relying primarily on terminal benchmark performance. A central task is therefore to identify which observables, or combinations of observables, can serve as state variables for describing how learning unfolds. Studying their evolution and relationships can yield reusable insights into how models learn.

3.2 nanoGPT Observable Library

Standard training metrics such as loss and aggregate parameter or gradient norms help monitor optimization, but do not fully describe learning dynamics. Smooth aggregate curves may conceal sharp changes within individual layers, while models with similar losses may follow different internal trajectories. **A systematic search for state variables therefore requires a broader space of candidate observables, measured throughout training and localized within the model**.

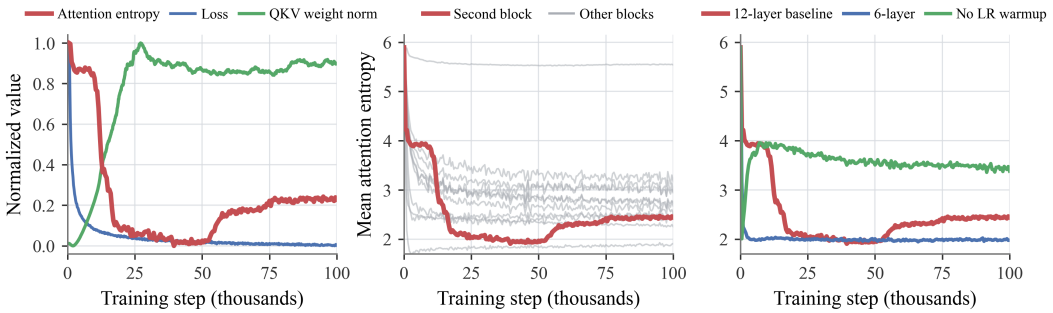


Figure 2: (a) Second-block mean attention entropy is compared with loss and QKV weight norm, with each curve independently normalized after the initial observation; entropy reveals a trajectory not visible in standard training metrics. (b) Mean attention entropy is compared across blocks; the second block follows a distinct trajectory. (c) Second-block entropy is compared across the 12-layer baseline, 6-layer, and no-warmup configurations; the transition may depend on model architecture and training configuration.

To support this search, we develop the nanoGPT Observable Library using Karpathy's standard GPT-2 Small training recipe (Karpathy, 2022; Radford et al., 2019) on OpenWebText (Gokaslan & Cohen, 2019). nanoGPT is compact, transparent, and easy to modify, while preserving the main components of a decoder-only Transformer (Vaswani et al., 2017). Instead of selecting a small number of metrics by hand, the library systematically combines model quantities, statistics, and architectural locations, and expands each measurement across compatible modules and layers. Combining weights, gradients, updates, optimizer states, activations, logits, and attention maps with different statistics produces a large, structured collection of observables. Their trajectories can then be compared across layers and experimental settings in a consistent way.

Among these automatically generated observables, we find a distinctive pattern in mean attention entropy. For each query token, we compute the entropy of its attention weights over key tokens and then average the result over queries, heads, and training examples. A higher value means that attention is more spread out, while a lower value means that it is more concentrated. In the 12-layer reference run, the second transformer block shows three stages: its entropy stays near 4, falls to a plateau near 2, and later rises toward 2.5. Other blocks do not show the same pattern. When we reduce the model to six layers, the entropy quickly falls to 2 and stays there. Removing learning-rate warmup changes the trajectory in another way: it falls briefly, rebounds toward 4, and then slowly declines. **These changes are not visible in the QKV weight norm, showing how the expanded observable space can reveal new questions about how models learn**.

These comparisons turn the attention-entropy pattern into specific research questions. Since the six-layer model does not show the same three-stage trajectory, is the transition related to network depth? Since removing learning-rate warmup makes the entropy rebound to a higher plateau, does warmup affect how broadly attention is distributed, or the effective context length used by the model? The library allows researchers to record such observations together with the relevant curve region, layer, training setup, and supporting comparisons, so that others can inspect the same evidence and test the resulting hypotheses. In this way, **the library provides a shared empirical basis for developing and testing hypotheses about how models learn**.

3.3 OPHIS

One may be interested about questions like "how observables can be useful for actual utillizations?". In this section, we will depict "OPHIS", a niche autoresearch paradigm for deep learning model insight scaling which aims to scale observables that results in actual model performance improvements. OPHIS stands for "observation", "problem", "hypothesis", "intervnetion" and speedup. Just like fluid dynamist, we let LLM agent to recognize the affecting factors (observables), list out the hypothesis, intervene the affecting factor, and to observe whether the hypothesis works as it proposed.

We tested the OPHIS looping paradigm on Karpathy's autoresearch training task (Karpathy, 2026), starting from Recursive's optimized solution (Recursive, 2026), to answer the following question: "Can observable regularization speedup the training of nanochat?". For a static machine learning task, observables can be effectively enumerated" since it is a static set. The goal of this task is to reach the lowest validation bits per bytes - reach the lowest entropy for maximum likelihood learning per unit of bytes, similar to Karparthy's autoresearch task.

OPHIS begins without relying on Internet knowledge or a library of previous experiments. Instead, it gives the agent rich context from roughly 6,000 internal observations. We changed the original five-minute challenge into a fixed 2,000-step challenge so that the agent could not exploit wall-clock-specific optimizations that would not generalize to training itself. All iterations ran on a single H200 GPU, and each candidate solution was evaluated ten times to distinguish real gains from kernel and training noise.

Under this protocol, the best OPHIS intervention reduced validation BPB by 7.43 baseline standard deviations relative to the re-evaluated RSI baseline (Figure 3). In a separate modular-addition grokking experiment, OPHIS also achieved higher substantial-improvement rates and lower failure rates than the LLM-only baselines. Across these two tasks, **OPHIS outperformed the evaluated LLM-based AutoResearch baselines.** Detailed experimental settings and results are provided in Appendix A.

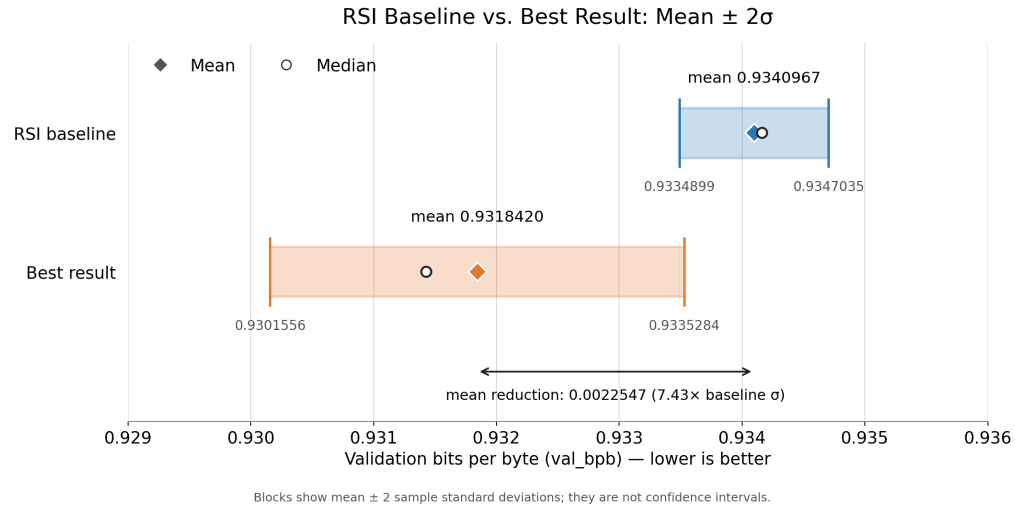


Figure 3: Validation BPB for the RSI best result and the best OPHIS result across repeated runs. Diamonds show means, circles show medians, and shaded ranges show the mean ± two sample standard deviations.

Beyond improving training performance, OPHIS can also reveal learning phenomena for further study. For example, the loop surfaced "forking", a failure mode involving rapid and severe overfitting early in training under certain architectural conditions (see Figure 5 in Appendix A.3). Systematically investigating such phenomena and accumulating the resulting insights may deepen our understanding of learning, help researchers avoid training failures, and guide further improvements in model performance.

3.4 COMFYRESEARCH

Scaling phenomena requires more than running a larger number of isolated experiments. It needs to unify the structure of experiments. To be more specific, it requires a way to construct, modify, compare, and share experiments. We introduce ComfyResearch, a visual environment in which machine-learning experiments are represented as executable and structured graphs.

Experiments as executable graphs. In ComfyResearch, an experiment is a directed graph whose nodes are computational primitives and whose edges represent data and control dependencies. A

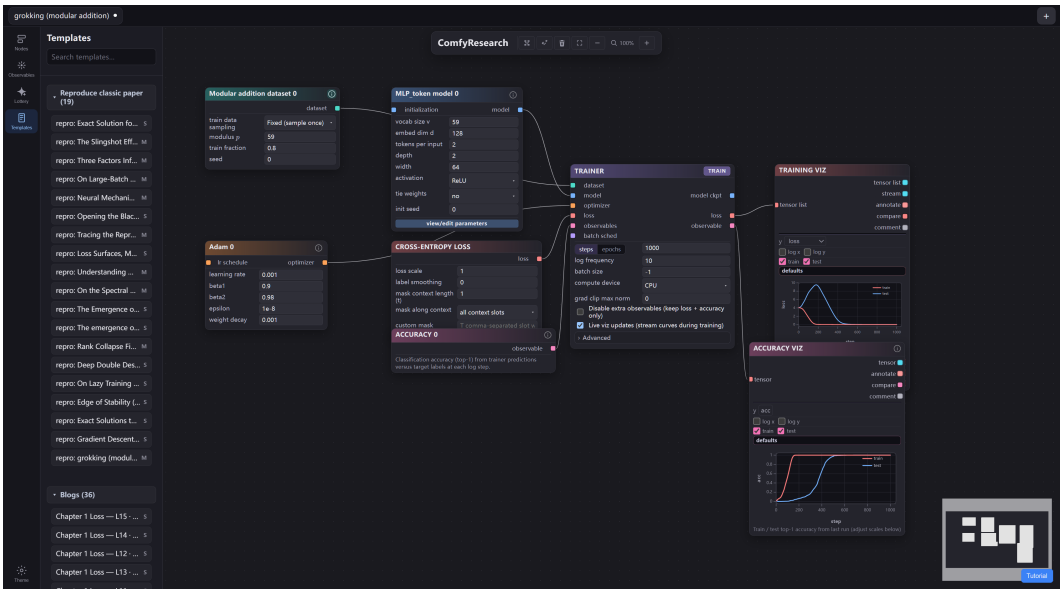


Figure 4: A ComfyResearch graph for a modular-addition grokking experiment. Dataset, token model, optimizer, loss, trainer, and accuracy/loss visualizations are composed into a single executable experiment, providing a concrete setting for inspecting the separation between memorization and generalization.

researcher composes datasets, models, optimizers, losses, trainers, observables, visualizations, and analysis nodes on a shared canvas. For example, a grokking experiment can be built by connecting a modular-addition dataset node to a small Transformer node, an AdamW optimizer node (Loshchilov & Hutter, 2019), a cross-entropy loss node, and a set of observable nodes that track test loss, accuracy(Figure 4). The graph formalism makes the experimental structure explicit: all necessary components are pre-built and the graph defines a unique experiment.

Reproducing classical experiments Besides flexible pre-defined components, ComfyResearch provides a library of blueprints and templates that reproduce established learning phenomena. Rather than writing a new training script for each study, researchers can reproduce classical phenomenon such as edge-of-stability training (Cohen et al., 2021), grokking (Power et al., 2022), or neural collapse (Papyan et al., 2020), and then vary individual nodes or hyperparameters to build upon these templates (Figure 6). These reusable starting points lower the cost of reproduction and make it easier to test whether an observation transfers to a new setting.

Generating new experiments through lottery Systematic exploration also requires generating experiments that have not been tried before. ComfyResearch includes a lottery mechanism that samples and perturbs graph configurations to discover new training regimes (Figure 7). By combining lottery generation with the observable library, researchers can scan a broad design space and identify configurations that produce interesting dynamics, and then generate new knowledge using observable and analysis nodes. For researchers that is not comfortable or familiar with the platform, lottery is also a good starting point and also useful for them, since it accelerates the discovery of new phenomenon.

4 Research Agenda

The explorations above provide starting points for scaling insight, while much work remains to be done. Further work should improve how we discover insights, accumulate and share them across studies, and use them to guide new research. We also need to understand when investing in insight is worthwhile under limited compute and time. We outline four research directions toward these goals.

4.1 IDENTIFYING OBSERVABLES FOR LEARNING DYNAMICS

We have also developed the Observable Library, a tool independent of the nanoGPT Observable Library that supports recording observables in a wider range of PyTorch training tasks. The tool is still at an early stage. A central challenge is to identify which observables, or combinations of them, can serve as effective state variables for learning dynamics. One direction is to systematically explore candidate observables, compare their trajectories, and test the resulting hypotheses through small, controlled experiments. A complementary direction is to use historical training data in a Bayesian framework to estimate which observable categories are likely to be useful for describing learning dynamics in different types of tasks. These posterior estimates could be updated as new training runs and observations become available, guiding which observables to examine next.

4.2 ACCUMULATING AND SHARING INSIGHTS

Scaling insight requires ways to accumulate knowledge across studies and share it beyond the researchers who produced it. One direction is to develop a domain-specific language (DSL) for expressing insights, including their claims and conditions of applicability. Such a language should make the insights understandable and reusable by both human researchers and AI systems. Further work should investigate how to test whether insights transfer across tasks and model scales. It should also examine how local regularities from different studies can be connected and how counterexamples can help refine their claims and scope. The goal is to let each study contribute knowledge that others can use, test, and extend.

4.3 TOWARDS NEXT-LEVEL AUTORESEARCH

OPHIS already uses an LLM agent to observe training, formulate hypotheses, and test interventions. Building on this early implementation, a key challenge is to improve agents' research taste, including their judgment about which questions and experiments are worth pursuing. One direction is to provide agents with accumulated insights and investigate how this knowledge can improve their research decisions. Another is to develop systems in which agents systematically observe learning dynamics, test hypotheses, and accumulate the resulting insights in a knowledge base. Agents should use new results to revise this knowledge and apply it in subsequent research. We call this proposed paradigm, in which agents both use existing insights and contribute new ones, the next-level AutoResearch.

4.4 MODELING THE VALUE AND COST OF INSIGHT

Acquiring insights also requires compute and time. Under a fixed budget, spending nearly all resources on additional observation and analysis may leave too little for training a model to the desired performance. This motivates modeling when insight-guided research can achieve better model performance than direct training under the same total compute and time constraints. The analysis should account for the costs of acquiring and testing insights, their reuse across tasks, and the fact that exploratory training can also produce useful models. We expect such modeling to yield a phase diagram showing which resource-allocation strategies are preferable under different conditions.

5 CONCLUSION

We argue that AI research should scale insight alongside compute, making reusable knowledge an explicit goal of experimentation. Insights need not provide complete mechanistic explanations to be useful, but they should be communicable and have a stated scope of applicability. Our explorations show how systematic observation can reveal learning phenomena, inform hypotheses, and guide interventions, providing starting points for both the science of AI and AI-for-AI research. How insights transfer across settings, and when the benefits of acquiring them justify their costs, remain open questions. Research should aim not only to improve the model under study, but also to produce knowledge that others can test, reuse, and extend.

AI USE STATEMENT

We used Codex, ChatGPT Work, and Claude Code to assist with manuscript drafting and language polishing, literature searches, figure preparation (including image generation and plotting code), and code development for the nanoGPT Observable Library, OPHIS, and ComfyResearch. The authors designed and led the research, including the OPHIS-based AutoResearch workflow. Within this workflow, Codex served as the research agent, inspecting training observations, formulating hypotheses, and implementing and testing interventions under the experimental protocol defined by the authors. The authors reviewed the AI-assisted work and experimental results, made the key research decisions, and take responsibility for the final content of this paper, including its claims, figures, and accompanying research artifacts.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkeStsCcKQ.

Anthropic. Learning more about Claude's mathematical capabilities. Anthropic research report, August 2026. URL https://www.anthropic.com/research/riemann-zeta. Published 10 August 2026; updated 13 August 2026. Accessed 6 September 2026.

Artur Avila, Manjul Bhargava, Caucher Birkar, Pierre Deligne, Yu Deng, Simon Donaldson, Hugo Duminil-Copin, Alessio Figalli, Martin Hairer, June Huh, Maxim Kontsevich, Elon Lindenstrauss, Pierre-Louis Lions, James Maynard, Curtis McMullen, Shigefumi Mori, Bao Chau Ngo, Andrei Okounkov, Peter Scholze, Stanislav Smirnov, Terence Tao, Maryna Viazovska, Cédric Villani, Wendelin Werner, and Efim Zelmanov. A severe misalignment of AI in mathematics. Joint declaration, September 2026. URL https://mathandai.org/.

Hasan Baig, Hailey Joren, and Joe Benton. TASTE: Can AI models judge AI safety research proposals? Anthropic Alignment Science Blog, August 2026. URL https://alignment.anthropic.com/2026/taste/. Published 28 August 2026. Accessed 7 September 2026.

Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027. URL https://doi.org/10.1017/S0962492921000027.

Stella Biderman, Mohammad Aflah Khan, Niloofar Mireshghallah, Catherine Arnett, Fazl Barez, and Naomi Saphra. Position: Don't just "fix it in post": A science of AI must study learning dynamics. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://openreview.net/forum?id=VbHh05xQ7y.

William H. Brock. *The History of Chemistry: A Very Short Introduction*. Oxford University Press, 2016. doi: 10.1093/actrade/9780198716488.001.0001. URL https://doi.org/10.1093/actrade/9780198716488.001.0001.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Nicholas Carlini. Building a C compiler with a team of parallel Claudes. Anthropic engineering blog, February 2026. URL https://www.anthropic.com/engineering/building-c-compiler. Published 5 February 2026. Accessed 7 September 2026.

Sadi Carnot. *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*. Bachelier, Paris, 1824. URL https://doi.org/10.3931/e-rara-9118.

Yueh-Han Chen, Jiaxin Wen, and Jan Hendrik Kirchner. Automated researchers can mitigate well-characterized alignment failures. *arXiv preprint arXiv:2608.28945*, 2026. doi: 10.48550/arXiv.2608.28945. URL https://arxiv.org/abs/2608.28945.

I. Bernard Cohen. *The Birth of a New Physics*. W. W. Norton & Company, revised and updated edition, 1985. URL https://wwnorton.co.uk/books/9780393300451-the-birth-of-a-new-physics.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.00065.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15594.

Henk W. de Regt and Dennis Dieks. A contextual approach to scientific understanding. *Synthese*, 144:137–170, 2005. doi: 10.1007/s11229-005-5000-4. URL https://doi.org/10.1007/s11229-005-5000-4.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2005.12729.

Andrew Ferguson, Marisa LaFleur, Lars Ruthotto, Jesse Thaler, Yuan-Sen Ting, Pratyush Tiwary, and Soledad Villar. The future of artificial intelligence and the mathematical and physical sciences (AI+MPS). *Machine Learning: Science and Technology*, 7(2):023001, 2026. doi: 10.1088/2632-2153/ae3e4e. URL https://doi.org/10.1088/2632-2153/ae3e4e.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus, 2019. URL https://skylion007.github.io/OpenWebTextCorpus/.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 558–567, June 2019. doi: 10.1109/CVPR.2019.00065. URL https://openaccess.thecvf.com/content_CVPR_2019/html/He_Bag_of_Tricks_for_Image_Classification_with_Convolutional_Neural_Networks_CVPR_2019_paper.html.

Liangze Jiang, Zachary Shinnick, Anton van den Hengel, Hemanth Saratchandran, and Damien Teney. Procedural pretraining: Warming up language models with abstract data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2601.21725.

Sara Kangaslahti, Elan Rosenfeld, and Naomi Saphra. Hidden breakthroughs in language model training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.15872.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Andrej Karpathy. nanoGPT. GitHub repository, 2022. URL https://github.com/karpathy/nanoGPT.

Andrej Karpathy. autoresearch. GitHub repository, 2026. URL https://github.com/karpathy/autoresearch.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025. URL https://arxiv.org/abs/2501.12599.

John Kounios and Mark Beeman. The cognitive neuroscience of insight. *Annual Review of Psychology*, 65:71–93, 2014. doi: 10.1146/annurev-psych-010213-115154. URL https://doi.org/10.1146/annurev-psych-010213-115154.

Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao, and Alán Aspuru-Guzik. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4:761–769, 2022. doi: 10.1038/s42254-022-00518-3. URL https://doi.org/10.1038/s42254-022-00518-3.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018. URL https://arxiv.org/abs/1807.03341. Presented at ICML 2018: The Debates.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34651–34663. Curran Associates, Inc., 2022. doi: 10.52202/068431-2511. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/dfc310e81992d2e4cedc09ac47eff13e-Paper-Conference.pdf.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01117.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://doi.org/10.1038/s41586-026-10265-5.

Thang Luong and Edward Lockhart. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. Google DeepMind research blog, July 2025. URL https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard
Published 21 July 2025. Accessed 7 September 2026.

Neil Rohit Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42834–42856. PMLR, 2025. URL https://proceedings.mlr.press/v267/mallinar25a.html.

Jialin Mao, Itay Griniasty, Han Kheng Teoh, Rahul Ramesh, Rubing Yang, Mark K. Transtrum, James P. Sethna, and Pratik Chaudhari. The training process of many deep networks explores the same low-dimensional manifold. *Proceedings of the National Academy of Sciences*, 121(12): e2310002121, 2024. doi: 10.1073/pnas.2310002121. URL https://doi.org/10.1073/pnas.2310002121.

Dmitri Mendeleev. The periodic law of the chemical elements. *Journal of the Chemical Society, Transactions*, 55:634–656, 1889. doi: 10.1039/CT8895500634. URL https://doi.org/10.1039/CT8895500634.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02292.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

OpenAI. Improved long gaps between primes. Research manuscript, 2026. URL https://cdn.openai.com/pdf/51126fac-1b68-4128-9666-c908bcc16033/long_gaps.pdf. Accompanying Lean formalization: https://github.com/openai/LongGapsBetweenPrimes. Accessed 6 September 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://doi.org/10.1073/pnas.2015509117.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Recursive. First steps toward automated AI research. Technical blog post, 2026. URL https://www.recursive.com/articles/first-steps-toward-automated-ai-research.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022. doi: 10.1109/IJCNN55064.2022.9891914. URL https://doi.org/10.1109/IJCNN55064.2022.9891914.

Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, Dhruva Karkada, Eric J. Michaud, Berkan Ottlik, and Joseph Turnbull. There will be a scientific theory of deep learning, 2026. URL https://arxiv.org/abs/2604.21691.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8df7c2e3c3c3be098ef7b382bd2c37ba-Abstract.html.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50528–50652. Curran Associates, Inc., 2024. doi: 10.52202/079017-1601. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The Clock and the Pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27223–27250. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56cbfbf49937a0873d451343ddc8c57d-Abstract-Conference.html.

A OPHIS

A.1 NANOCHAT RESULTS

We replaced the original five-minute budget with a fixed 2,000-step training budget. All runs used a single H200 GPU, and each candidate solution was evaluated ten times. Under this protocol, we re-evaluated Recursive's optimized solution (Recursive, 2026), obtaining a baseline mean validation BPB of 0.9340967. Using only 13% of nanochat's available observables, OPHIS identified 376 interventions that improved on this baseline by at least three baseline standard deviations. The best intervention achieved a mean validation BPB of 0.9318420, a reduction of 0.0022547, equivalent to 7.43 baseline standard deviations. Lower BPB is better. Figure 3 in the main text shows the repeated-run comparison.

A.2 GROKKING RESULTS

The unmodified baseline reached 95% test accuracy at approximately 1,250 training steps. The LLM-only baseline received the task, architecture, optimizer, and training configuration, but not OPHIS measurements, ranked internal variables, or previously discovered interventions. Of its 100 proposed candidates, 76 were implemented and tested; 24 were excluded, unsupported, or outside the comparison setting. Table 1 reports outcomes among tested interventions, including LLM proposals in the OPHIS-like intervention subspace. Substantial improvements reached the target accuracy clearly earlier than the unmodified baseline; marginal gains were not counted.

Table 1: Intervention outcomes on modular-addition grokking. Percentages use the number of tested interventions in each row as the denominator.

Method	Tested	Substantial improvement	Failure
OPHIS	350	255 (72.9%)	48 (13.7%)
LLM, full baseline	76	44 (57.9%)	32 (42.1%)
LLM, OPHIS-like subspace	30	21 (70.0%)	9 (30.0%)

A.3 FORKING

Figure 5 illustrates "forking," the rapid onset of severe overfitting observed under certain architectural conditions during our OPHIS experiments. This example shows how experiments aimed at improving model performance can also reveal learning phenomena for further investigation.

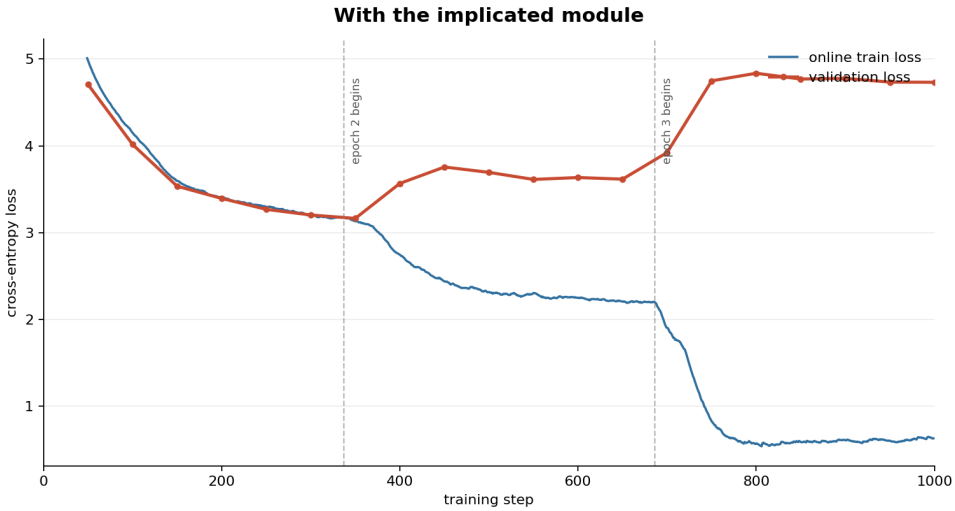


Figure 5: With the implicated module active, training loss rapidly separates from validation loss early in replay, a phenomenon we call "forking".

B ADDITIONAL COMFYRESEARCH EXAMPLES

Figure 6 shows an edge-of-stability template in ComfyResearch, referenced in Section 3.4. A deterministic synthetic regression dataset, an MLP, SGD, exact Hessian eigenvalue tracking, and training visualizations are pre-connected, so researchers can reproduce the phenomenon and then vary individual components.

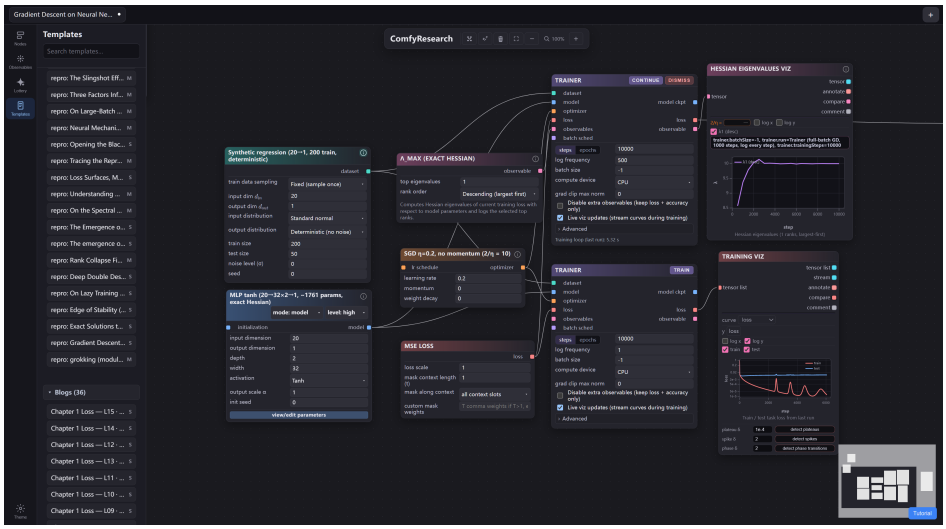


Figure 6: A ComfyResearch template for an edge-of-stability experiment. A deterministic synthetic regression dataset, an MLP, SGD, exact Hessian eigenvalue tracking, and training visualizations are represented as connected components, making the relationship between curvature and optimization dynamics directly inspectable.

Figure 7 shows the Lottery interface in ComfyResearch, referenced in Section 3.4. The interface connects a numerical or textual seed to a concrete generated experiment and exposes the resulting dynamics for further inspection and modification.

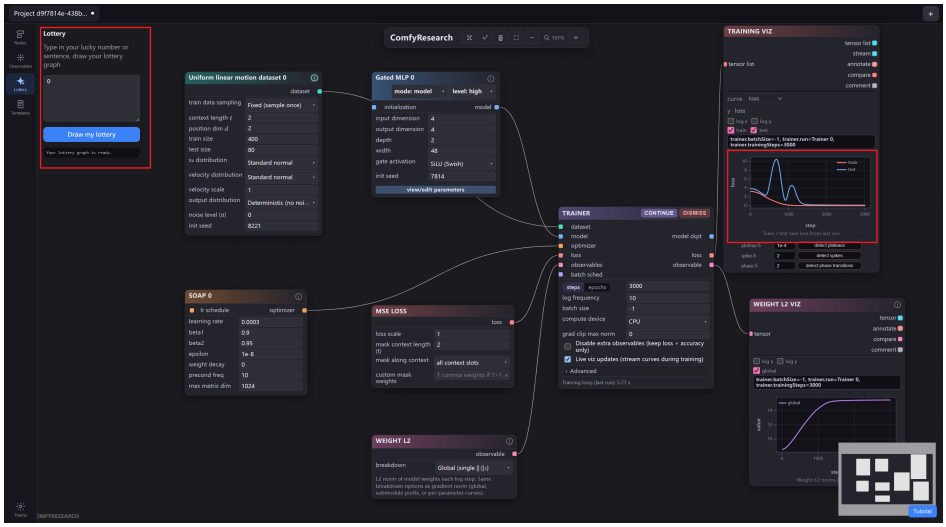


Figure 7: The Lottery interface in ComfyResearch, shown with the generated graph and its resulting training curves highlighted. The interface connects a numerical or textual seed to a concrete experiment and exposes the resulting dynamics for further inspection and modification.